

LA-UR-18-27414

Approved for public release; distribution is unlimited.

Title: Injecting Systematic Faults to Evaluate Risks of a Multi-Cluster Slurm Database

Author(s): Dobson, Nicole Anne
Seamons, Calvin Daniel
Morrell, Zachary Alexander

Intended for: HPC Mini-Showcase, ISTI Day

Issued: 2018-08-03

Disclaimer:

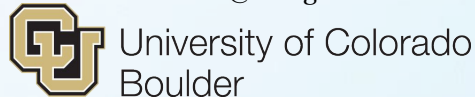
Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the Los Alamos National Security, LLC for the National Nuclear Security Administration of the U.S. Department of Energy under contract DE-AC52-06NA25396. By approving this article, the publisher recognizes that the U.S. Government retains nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

Injecting Systematic Faults to Evaluate Risks of a Multi-Cluster Slurm Database

Calvin Seamons
calvindseamons@gmail.com



Nicole Dobson
ndobson@lanl.gov



Zachary Morrell
zmorrell@unm.edu



Outline

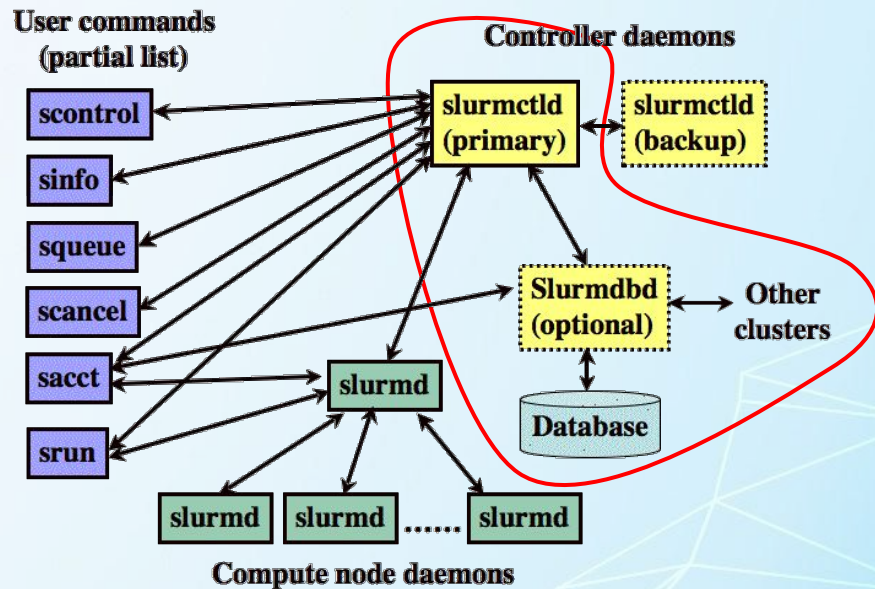
- Background
- Methodology
 - Software Configuration
 - Fault Injections
 - Data
- Potential Mitigations
 - Monitoring Log Messages
- Conclusion & Future Work

Background

What is Slurm?



Basic Functionality



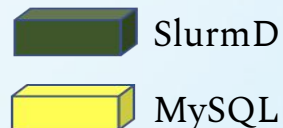
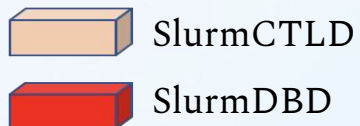
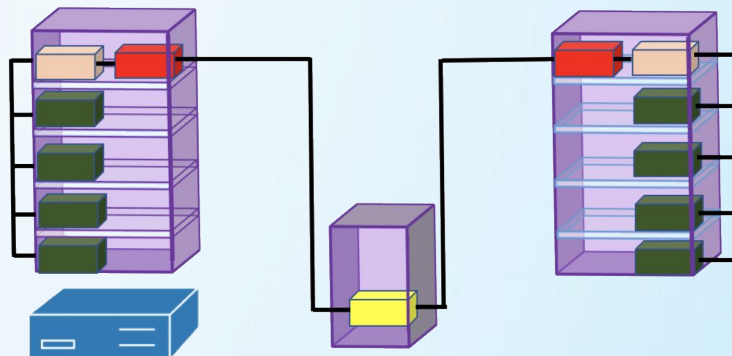
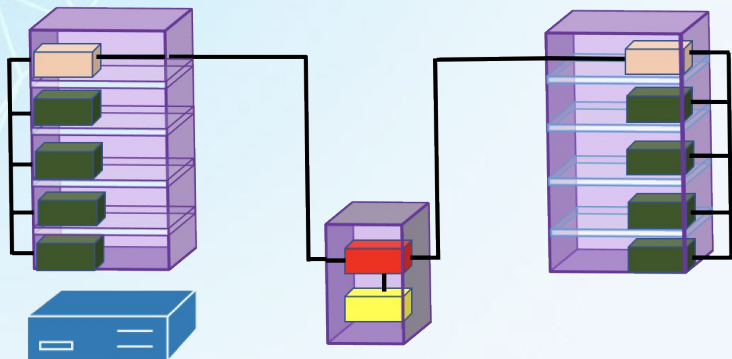
Slurm Database

Benefits of a Multi-Cluster Slurm Database

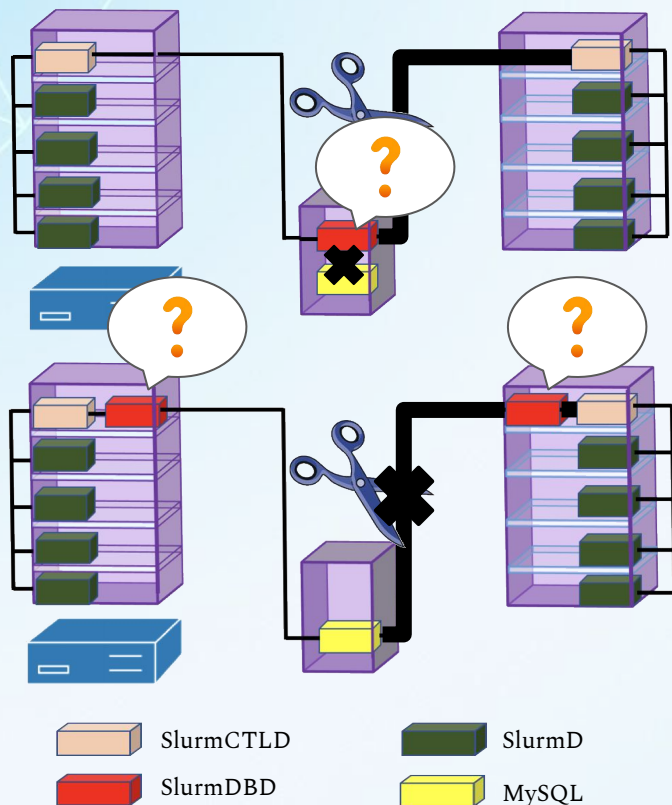
- Comparative analysis
 - Start Time, End Time, Project, Energy Consumption, JobID, Elapsed Time
- Minimizes user data redundancy
- Monitoring and profiling of users
 - Monitoring focused on a single system
- Security benefits



Software Configuration



Fault Injections



Baseline

- 200 jobs running HPL configured to run for approximately 20 sec each

Network Fault Injection

- Unplugged ethernet cable

Transaction Fault Injection

- Database does not listen to SlurmDBD

High Influx of Jobs

- Running hostname on all processors on all compute nodes

What We Expect To See



If the fault is catastrophic

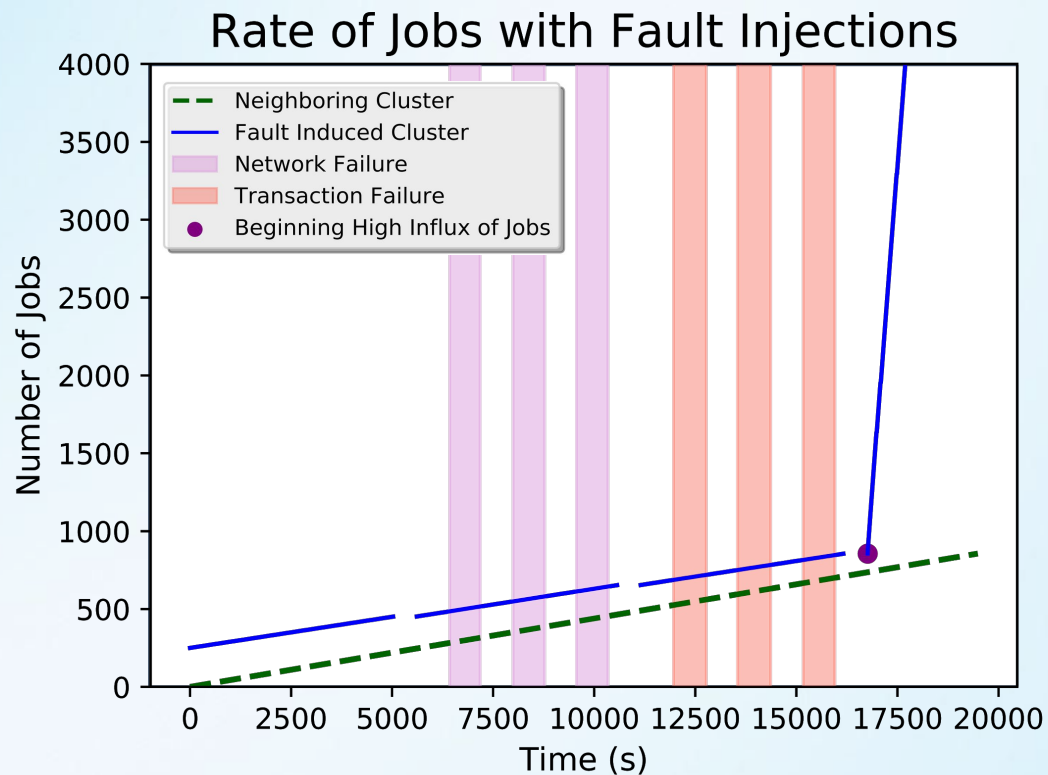
- Non-linear lines = performance degradation
- Gaps in the lines = data loss
- Line plateaus = Slurm stops jobs



If the fault does not cause problems

- Linear lines
- No gaps (except between tests)
- Lines continuous

Fault Injections



**Takeaway:
the lines
are linear!!**



Fault Injections: Limiting Partition Sizes



If the fault is catastrophic

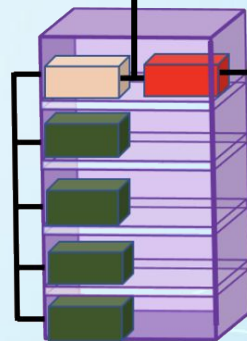
- Non-linear lines = performance degradation
- Gaps in the lines = data loss
- Line stops = Slurm stops jobs



If the fault does not cause problems

- Linear lines
- No gaps (except between tests)
- Lines continuous

Spool



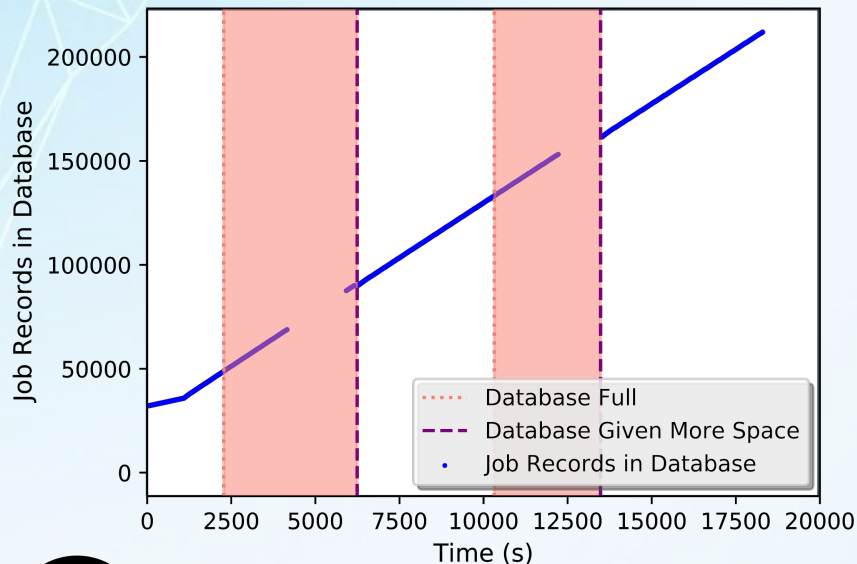
SlurmCTLD
SlurmDBD



SlurmD
MySQL

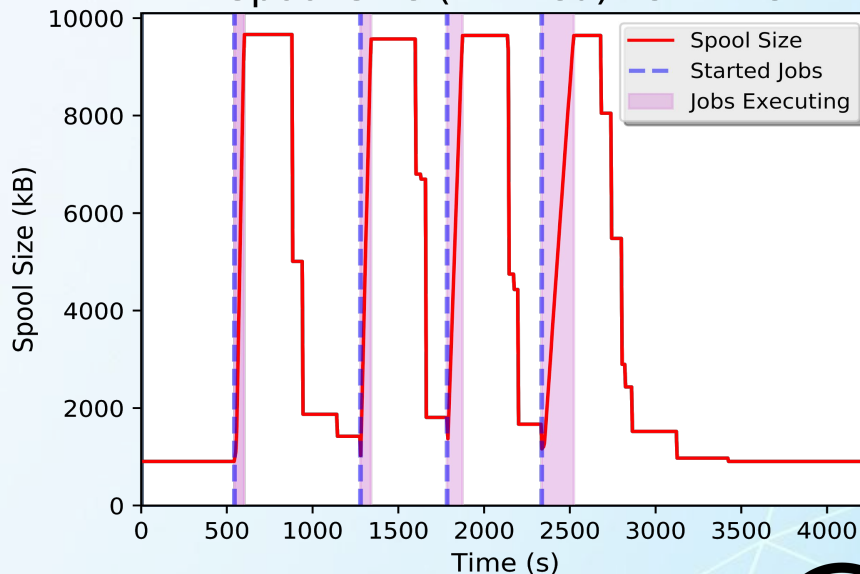
Fault Injections: Limiting Partition Sizes

Data Loss When Database is Full



1) Records of jobs are lost and

Spool Size (Limited) vs. Time

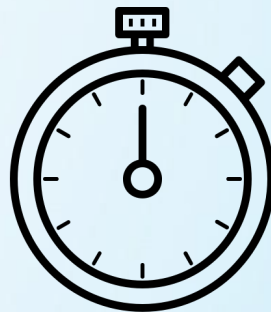


Faults manifest in two ways:

2) Job submission stoppage

Time to Repair and Potential Mitigations

- Provide ample capacity for the database
- Set up a monitoring process for the size of the MySQL database and the spool
- When a network connection failure occurs or the database runs out of disk space, the compute nodes can be set to drain to prevent the spool being overloaded



Monitoring Log Messages

SlurmCTLD log messages when MySQL database is full

```
[2018-07-23T09:03:52.101] error: It looks like the storage has gone away trying to reconnect  
[2018-07-23T09:03:52.101] error: mysql_real_connect failed: 2003 Can't connect to MySQL server on '192.168.1.253' (111)  
[2018-07-23T09:03:52.101] error: unable to re-connect to as_mysql database  
[2018-07-23T09:03:52.101] fatal: You haven't init'd this storage yet.
```

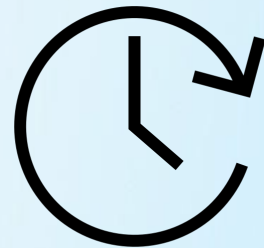
Command `sacct` from Master node when MySQL database is full

```
sacct: error: slurm_persist_conn_open_without_init: failed to open persistent connection to localhost:6819: Connection refused  
sacct: error: slurmdbd: Sending PersistInit msg: Connection refused  
sacct: error: Problem talking to the database: Connection refused
```

Results

- No significant performance difference for where SlurmDBD is located
 - Easier to have the singleton SlurmDBD rather than multiple, also more secure
- Spool on master nodes handles job account information when errors occur
 - Spool can load-shed, appears to have some internal Slurm regulation
- No predictive log messages found
 - Log messages thrown upon failure
- Repair time is reasonable compared to the time until data loss
 - Safe downtime depends on system and job load

Future Work



- Vary parameters
 - Test with nonidentical clusters and more than 2 clusters
 - Use a more diverse/realistic suite of jobs
- Vary database implementation
 - Different database “drop in”
 - Limit swap space
 - Model the spool capacity with greater precision to reduce down time
- Experiment with different monitoring techniques to enhance early detection

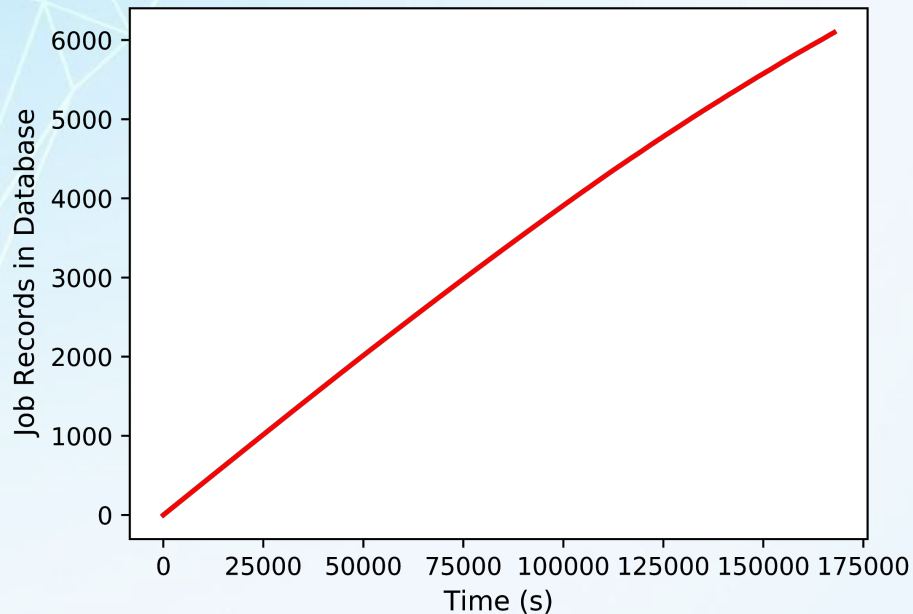
Questions?

Acknowledgements

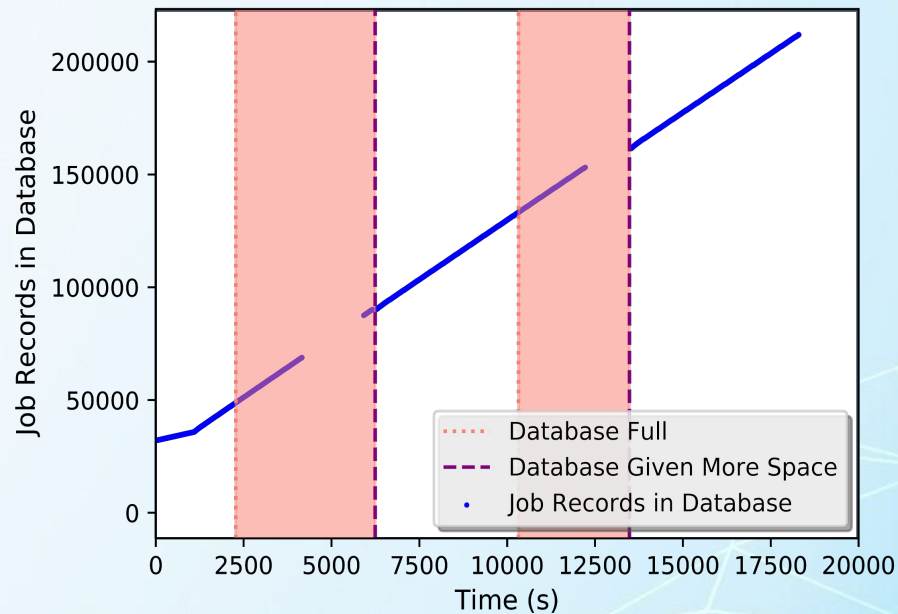
Mentors: Steven Senator, Joshi Fullop, Hai Ah Nam, Monique Morin, Lena M. Lopatina
Director: Alfred Torrez
Instructors: Hunter Easterday, Colette Caskie

Time to Losing Data Dependent on Jobs

Data Stored When Database Not Reachable



Data Loss When Database is Full



What Information is Stored in the Database

Fields available:

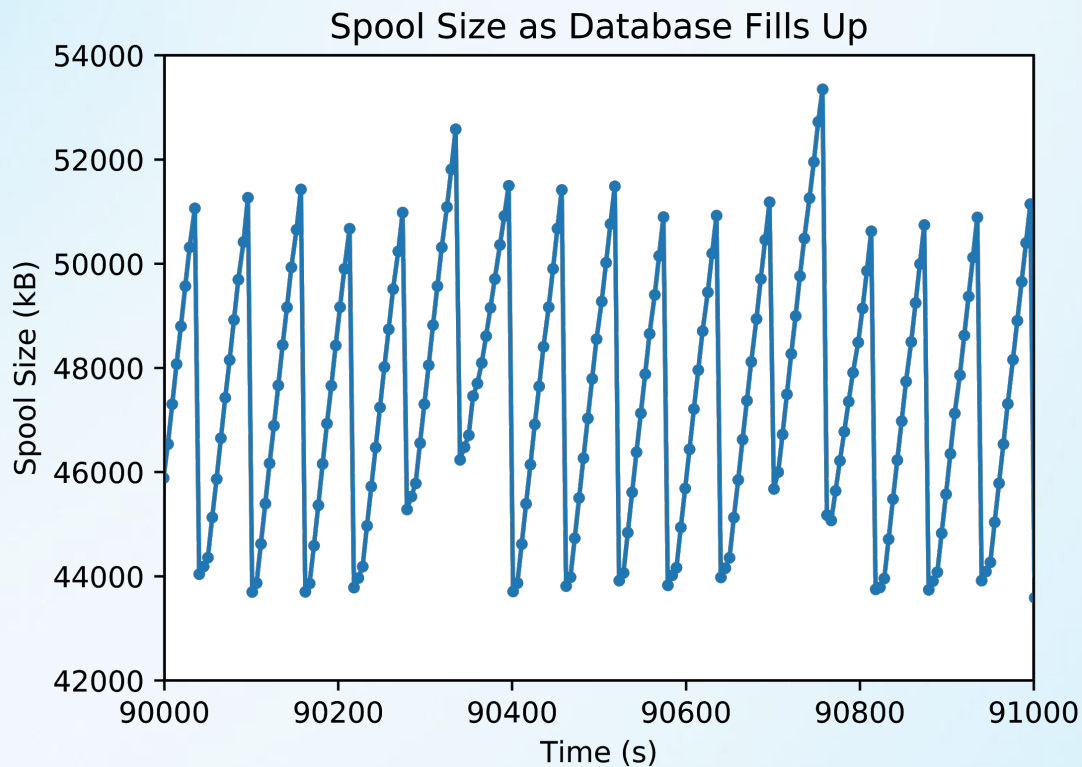
Account	AdminComment	AllocCPUS	AllocGRES
AllocNodes	AllocTRES	AssocID	AveCPU
AveCPUFreq	AveDiskRead	AveDiskWrite	AvePages
AveRSS	AveVMSize	BlockID	Cluster
Comment	ConsumedEnergy	ConsumedEnergyRaw	CPUTime
CPUTimeRAW	DerivedExitCode	Elapsed	ElapsedRaw
Eligible	End	ExitCode	GID
Group	JobID	JobIDRaw	JobName
Layout	MaxDiskRead	MaxDiskReadNode	MaxDiskReadTask
MaxDiskWrite	MaxDiskWriteNode	MaxDiskWriteTask	MaxPages
MaxPagesNode	MaxPagesTask	MaxRSS	MaxRSSNode
MaxRSSTask	MaxVMSize	MaxVMSizeNode	MaxVMSizeTask
McsLabel	MinCPU	MinCPUNode	MinCPUTask
NCPUS	NNodes	NodeList	NTasks
Priority	Partition	QOS	QOSRAW
ReqCPUFreq	ReqCPUFreqMin	ReqCPUFreqMax	ReqCPUFreqGov
ReqCPUS	ReqGRES	ReqMem	ReqNodes
ReqTRES	Reservation	ReservationId	Reserved
ResvCPU	ResvCPURAW	Start	State
Submit	Suspended	SystemCPU	Timelimit
TotalCPU	UID	User	UserCPU
WCKey	WCKeyID	WorkDir	

Spool Placed on Smaller Partition

Error in output file

```
sbatch: error: Batch job submission failed: I/O error writing script/environment to file
```

Spool Size Load Shedding



Network Topology

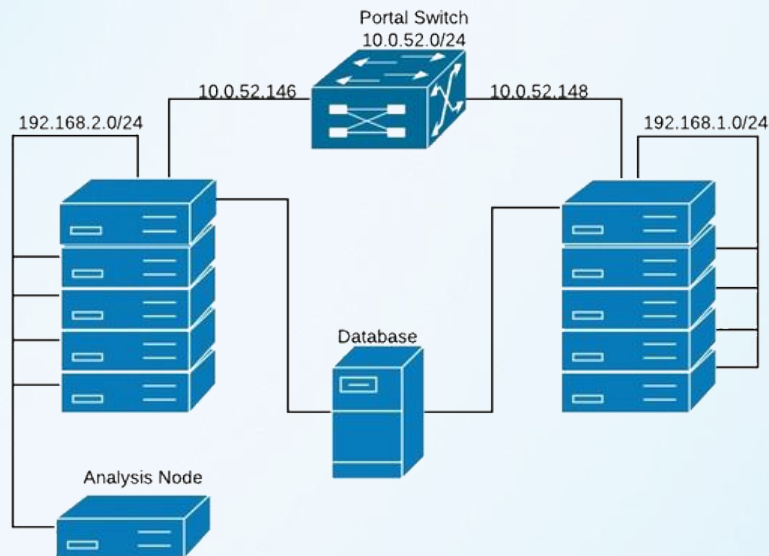


Image Citations

Created by Yuri Mazursky
from Noun Project

Created by Setyo Ari Wibowo
from Noun Project

Created by Dmitriy Ivanov
from Noun Project

Created by Sergio Barros
from Noun Project

Created by Made x Made
from Noun Project